A digital approach to protein identification and quantification using tandem nanopores and peptidases and search through a proteome database

G. Sampath

Abstract. A single-molecule method of identifying proteins based on electrical measurements and database search without labels or immobilization is considered. It uses electrolytic cells with two or three nanopores in tandem and one or two peptidases covalently attached to the *trans* side of a pore. An unknown protein is digested into peptides ending in a known amino acid; the peptides enter the cell, pass through the first pore, and are fragmented by a high-specificity endopeptidase. The second enzyme, if present, is an exopeptidase that cleaves the fragments into residues after the second pore. Level transitions in a blockade pulse due to the pore ionic current or transverse current pulse caused by a fragment in the second pore or individual such pulses caused by single residues in the third pore are counted. N residue-specific cells produce N integer lists from which a partial sequence is assembled. Search through the Uniprot database shows that for small N (3 to 5) over 98% of proteins in the human proteome can be identified from such sequences. A Fokker-Planck model is used to derive minimum enzyme turnover intervals required for correct sequencing. With thick (80-100 nm) pores the pulse width is ~1 μs/residue, which is within the capability of CMOS detector circuits. If digested peptides are assumed to enter a cell in random order then over a long run the quantity of a protein in a mixture of proteins can be estimated from the number of its identifying peptides.

Keywords: protein identification and quantification; peptide sequencing; electrolytic cell; nanopore; human proteome

1 Introduction

Unlike genome sequencing, which is largely aimed at extracting bio-markers such as gene mutations indicative of risk to diseases like cancer and diabetes, medical diagnostic procedures for patient treatment and care focus on the identification of cell constituents such as proteins. Often the identity and quantity of a protein in an assay are more useful than the sequence. While whole genome sequencing[1] has advanced rapidly with the emergence of several new techniques, sequencing and identification of proteins are largely based on the established techniques of Edman degradation,[2] gel electrophoresis,[2] and mass spectrometry.[3] Whether genome or protein, sequencing is based on bulky or expensive devices and/or time-consuming procedures; this has led to efforts aimed at developing portable/hand-held low-cost fast-turnaround devices.[4,5] In particular, nanopores have been investigated for their potential use in the analysis and study of DNA[6] and proteins/peptides.[7-10] Recently a tandem electrolytic cell with cleaving enzymes was proposed for sequencing of DNA[11] and peptides.[12] It has two single cells in tandem, with the structure [*cis*1, upstream pore (UNP), *tran*s1/*cis*2, downstream pore (DNP), *trans*2]. An enzyme covalently attached to the downstream side of UNP successively cleaves the leading monomer in a polymer threaded through UNP; the monomer translocates through DNP where the ionic current blockade it causes is used (along with other discriminators[12]) to identify it. With DNA the enzyme is an exonuclease,[11] with peptides it is an amino or carboxy peptidase.[12] The process is label-free and does not require immobilization of the analyte.

Here a low-cost alternative to conventional methods for protein identification is proposed in which a partial sequence is obtained for a peptide and used to identify the protein. Sequencing is based on a single tandem cell[11,12] and an endopeptidase (*Method 1*) or a double tandem cell, endopeptidase, and exopeptidase (*Method 2*). The first enzyme breaks the peptide into fragments, the second breaks fragments into residues. The fragments/residues translocate through a pore and cause ionic current blockades or modulate a transverse current across the pore;[6] the pore/transverse current pulses or level transitions within are counted. In both methods, N (~3 to 5) tandem cells, each with an endopeptidase specific to a different amino acid, produce N lists of integers corresponding to the positions of the amino acid in the peptide sequence, from which a simple algorithm assembles a partial sequence. The protein is then identified by comparing the latter with sequences in a protein database. With a mixture of proteins the quantity of a protein in the mixture is estimated from the number of identifying peptides. The approach may be extended to cover modified amino acids.

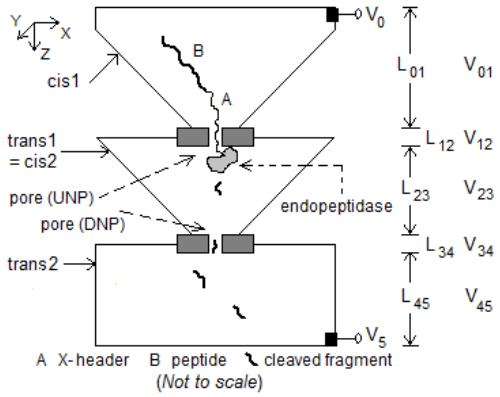
This is a digital technique based on pulse counting, it differs from other nanopore sequencing and identification techniques based on analog measurements of pulse magnitude or width (equivalently analyte residence time in a pore) in a pore ionic or transverse current.[6] The sequencing aspect is reminiscent of the Maxam-Gilbert[13] and Sanger[14] methods for DNA, wherein independent channels are used for A, T, C, and G, and subsequences are separated by length and terminal base type. The identification aspect resembles database-centered methods used in mass spectrometry such as Peptide Sequence Tags.[3] The approach is similar in some ways to a recent theoretical proposal[15] in which fluorescent labels specific to a set of amino acids are attached to residues in a peptide immobilized on a glass substrate. The labels are optically detected when the N-terminal residues are removed one after the other in a series of Edman degradation cycles. The optical output is used to partially sequence a peptide and then identify it in a proteome. In contrast, the proposal presented here does not require analyte immobilization, labeling, or repeated wash cycles.

2 Protein identification and quantification: method and materials

An unknown protein P_x is identified in six stages:

1. Fragment copy of P_x into peptides ending in amino acid X_0.
2. Break peptide copy into fragments ending in amino acid $X_1 \neq X_0$ (*Methods 1* and *2*). Break a fragment into individual residues (*Method 2*).
3. Find number of residues in fragment obtained in Stage 2.
4. Repeat Stages 1 through 3 for other amino acids X_2, X_3, ...
5. Assemble partial sequence from length information obtained in Stage 3. Mark unknown residues with wild card *.
6. Match partial sequence with sequences in proteome of interest and identify P_x (hopefully uniquely).

Stage 1. A highly specific chemical or peptidase is used. Examples include cyanogen bromide, which cleaves after methionine (M), and GluC protease, which cleaves after glutamic acid (E). Both cleave on the C terminal side and result in fragments ending in M or E respectively. More such agents/peptidases are available and are listed in Table A-4 in the Appendix. (See online review,[16] from which the table has been adapted, for a list of comprehensive references.)
Stage 2. The positions of occurrence of a residue in a peptide are obtained by targeting it with a highly specific peptidase in a tandem cell. Peptidases with high specificity include GluC (E), ArgC proteinase (arginine R), AspN endopeptidase (aspartic acid D), and LysC lysyl endopeptidase (lysine K). Others with high specificity but some ambiguity include serine proteinase (E or D) and neutrophil elastase (valine V or alanine A).[16] Similarly a peptide fragment can be cleaved into individual residues by an exopeptidase capable of cleaving a wide range of residue types at the carboxyl or amino end. Examples include Carboxypeptidase I (CPD-Y), Carboxypeptidase II (CPD-M-II), and Leucine Aminopeptidase (LAP).[17-19]
Stage 3. A tandem cell is used to count level transitions in a pore ionic current or transverse current pulse that is modulated by a fragment translocating through a nanopore or individual such pulses due to cleaved residues. Two methods are available.



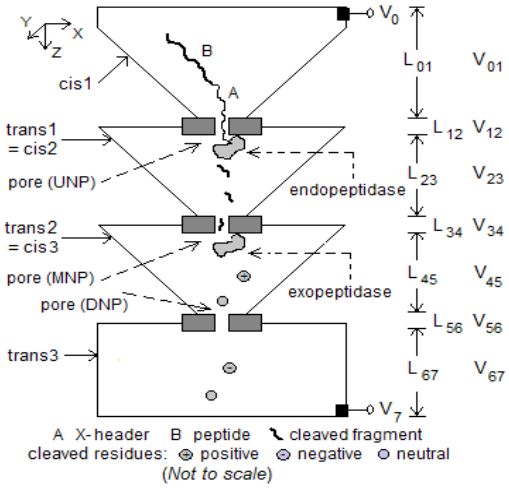


Figure 1 (left). Tandem cell for peptide sequencing. Dimensions: 1) *cis*1: box of height 1 μm tapering to 100 nm^2; 2) membrane with UNP: length 10-20 nm, diameter 10 nm; 3) *trans*1/*cis*2: box of height 1 μm tapering from 1 μm^2 to 10 nm^2; 4) membrane with DNP: length 10-20 nm, diameter 3 nm; 5) *trans*2: box of height 1 μm, side 1 μm. Endopeptidase covalently attached to downstream side of UNP. Electrodes at top of *cis*1 and bottom of *trans*2. V_{05} = ~115 mV.
Figure 2 (right). Double tandem cell for peptide sequencing. Dimensions: 1) *cis*1: box of height 1 μm tapering to 100 nm^2; 2) membrane with UNP: length 10-20 nm, diameter 10 nm; 3) *trans*1/*cis*2: box of height 1 μm tapering from 1 μm^2 to 10 nm^2; 4) membrane with MNP (middle nanopore): length 10-20 nm, diameter 3 nm; 5) *trans*2/*cis*3: box height 1 μm tapering from 1 μm^2 to 10 nm^2; 6) membrane with DNP: length 40-50 nm, diameter 3 nm; 7) *trans*3: box of height 1 μm, cross-section 1 μm^2. Endopeptidase (exopeptidase) covalently attached to downstream side of UNP (MNP). Electrodes at top of *cis*1 and bottom of *trans*3. V_{07} = ~180 mV.

In *Method 1* the structure in Figure 1 is used. A peptide with a poly-X header (X = one of the charged amino acids: Arg, Lys, Glu, Asp; the charge on X depends on the pH value) is drawn into UNP by the electric field due to V_{05} (= ~110 mV), most of which (~98%) drops across the two pores.[6] An endopeptidase specific to amino acid AA attached downstream of UNP cleaves the peptide after (or before) all n (≥ 0) points where AA occurs. The resulting n+1 fragments translocate to and through DNP, where level crossings in the resulting pore ionic current blockade or a transverse current across DNP may be used to count the residues in a fragment.

In *Method 2* the double tandem cell in Figure 2 is used. A peptide is cleaved into fragments by an endopeptidase as in *Method 1*. An exopeptidase (amino or carboxy) covalently attached downstream of the middle nanopore (MNP) cleaves residues from a fragment; the residues translocate through DNP and blockade the pore ionic current or modulate the transverse current. The resulting single pulses are counted.

In both methods each tandem cell specific to an amino acid produces an ordered list of integers equal to the lengths of successive fragments in which the last residue is the target. If a cell generates a single integer, the target is not in the peptide.

Stage 5. The peptide is partially assembled using the following procedure:

- Replace fragment lengths from cell with cumulative lengths (= target positions in peptide) and target identities.
- Invert position-identity pairs.
- Merge resulting sequences.
- Insert wild card * in all other positions in sequence.

Stage 6. Standard string matching algorithms can be used to search for the partial sequence among the set of sequences in a protein database such as Uniprot[20] or PDB. More general matching algorithms[21] may be used if desired.

2.1 Database search and results

The number of identifying sub-sequences per protein was computed for the following set of cleavage choices: M (first stage); R, K, D, E (second stage). An exhaustive search of sequences in the human proteome (Uniprot database id UP000005640, manually reviewed subset with 20207 sequences) was done. Computation is in four steps:

1. All subsequences ending in M are extracted, they correspond to the peptides generated by the action of cyanogen bromide (see above). Every one of these peptides has exactly one M which is also the last residue in the peptide.
2. In four individual cells specific to R, K, D, or E a copy of each peptide is cleaved after every occurrence of the target (R, K, D, or E). The resulting subsequences are the fragments generated by ArgC, LysC, AspN, or GluC respectively.
3. The partial peptide sequence is assembled from these fragments using the algorithm in Stage 5 above.
4. To find out if a peptide is a unique identifier the wild card * is entered into every position in the peptide sequence where R, K, D, E, and M do not occur. The resulting string is then matched with every other peptide (similarly filled with *).

Example: Consider the protein P31946 in the human proteome (Uniprot id UP000005640).[20] The following is one of three peptides that uniquely identify it in the proteome: KAVTEQGHELSNEERNLLSVAYKNVVGARRSSWRVISS IEQKTERNEKKQQM. With cells targeting R, K, D, and E, the corresponding length lists are R:{15, 14, 1, 4, 11}, K:{1, 22, 19, 6, 1}, D:{}, E:{5, 4, 1, 2, 26, 4, 3}, and M:{52}. The corresponding position lists are R:{15, 29, 30, 34, 45}, K:{1, 23, 42, 48, 49}, D:{}, E:{5, 9, 13, 14, 40, 44, 47}, and M:{52}, where the position value is obtained as the sum of all lengths in the list up to and including the current one. Inverting and merging leads to the partial sequence K***E***E***EER******K*****RR***R*****E***ER*EKK**M.

The percentage of proteins with at least one identifying subsequence (created by cyanogen bromide) is found to be 97.8%. The number of identifiable proteins can be increased by sequencing a protein with other combinations of cleaving chemical/enzyme in Stage 1 and peptidases in Stage 2. For example, instead of cyanogen bromide targeting M, GluC can be used to generate the set of peptides in the protein that end in E. Another possibility is to use diazonium to cleave after Tyr (Y) in the first stage. Various combinations of peptidases can be considered for the second stage; see Section A-6 in the Appendix. Figure 3 shows the distribution of the number of proteins vs the number of identifying peptides in a protein for two sets of cleavage choices in Stages 1 and 2.

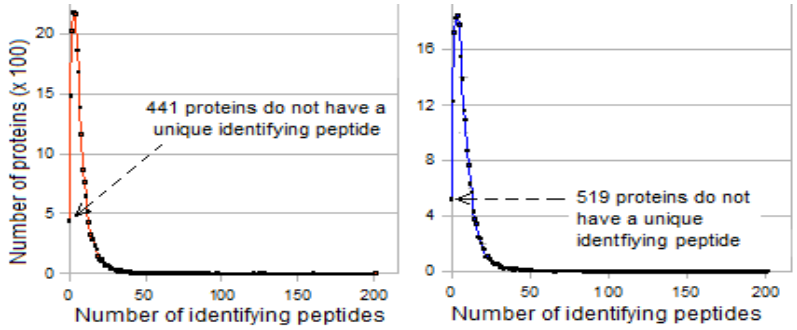


Figure 3. Number of proteins with a given number of identifying peptides in the human proteome (UP000005640), manually reviewed subset with 20207 sequences: (a) cleave M in stage 1 and R, K, D, E in stage 2; (b) cleave Y in stage 1 and R, K, D, E in stage 2

Table 1: Shows increase in number of identifiable proteins from multiple runs with different cleavage targets

Run no.	Cleaved target in first stage	Targeted in second stage	Identified proteins / not identified	Not identified in this run but identified in the other	Total unique proteins identified in both runs / not identified
1	M	R, K, D, E	19766 / 441	69	19835 / 372
2	Y	R, K, D, E	19688 / 519	147	

The total coverage is the union of the sets of proteins with at least one unique identifier obtained from all these

combinations. Table 1 shows the increase when sequencing is done twice by targeting M or Y in the first stage; both times R, K, D, and E are targeted in the second stage. With enough combinations 100% coverage may be possible.

2.2 Quantifying a protein in a mixture

Consider an assay with a mixture of proteins. The output of Stage 1 is the set of all the peptides from all the proteins in the mixture; they are the result of the cleaving action of the chemical agent or peptidase used. On input to Stage 2, the peptides enter a cell (which is designed to cleave after a given amino acid such as R, K, D, or E) in some random order; the partial sequences obtained are used to identify the container protein as described earlier. Consider a mixture $\{ (N_i, P_i, I_i): i = 1, 2, ... \}$ where N_i is the number of molecules of the i-th protein in the mixture, P_i the number of peptides per molecule of the protein (this is equal to the number of peptides created in Stage 1 from a single molecule), and I_i ($0 \le I_i \le P_i$) the number of identifying peptides per molecule. For a given chemical agent/peptidase in Stage 1 the P_is are known, and for the set of peptidases used in cells in Stage 2 the I_is are known by computation. N_i is the desired unknown. For example, with M targeted in Stage 1 and R, K, E, and D targeted in Stage 2, for the protein P31946 in the human proteome $P_i = 9$ and $I_i = 3$. Peptides generated in Stage 1 from the mixture enter a cell in succession (in some random order) and are partially sequenced and identified. Let the number of peptides in protein i that have been identified in the run so far be $I_{i\text{-measured}}$. If peptide entry into a cell is totally random, then after a sufficiently long run N_i can be estimated as

$$\tilde{N}_i \approx I_{i\text{-measured}} / I_i \qquad (1)$$

If the total number of peptides processed in the run is N_{total} then the number of peptides that do not yield identifying information is

$$N_{non\text{-identifying}} = N_{total} - \sum I_{i\text{-measured}} \qquad (2)$$

where the summation is over all the identified proteins. This number includes peptides that are found in more than one protein and may also include impurities in the assay sample. If the sample is not pure there seems to be no easy way to separate the two so unidentified sample proteins that are not impurities remain unestimated (even though their percentage is likely to be small).

3 Necessary conditions for correct sequencing

Nanopore-based sequencing relies on the ability to measure changes in current flow when an analyte molecule is present. This current may be an ionic current from *cis* to *trans*, a transverse electronic current across the pore membrane, or a transverse tunneling current across a gap in the membrane.[6] The measurement ability is closely related to the bandwidth of the detector, see discussion in Section 4 below.

Since the charge carried by a peptide is highly variable and may be negative, 0, or positive, the two methods described above rely on diffusion as the primary mechanism for translocation of a fragment or residue, modified by the drift field. They are studied through the properties of the basic tandem cell, which has been modeled with a Fokker-Planck equation.[11,12] Central to the model is the solution of a boundary value problem in which the *trans* side of a pore is viewed as a reflecting boundary for a cleaved fragment or residue, so the net diffusion tends to be in the *cis*-to-*trans* direction (with V_{05}, $V_{07} > 0$). The main quantities of interest are the mean E(T) and variance $\sigma^2(T)$ of the time T taken by a particle to translocate through a *trans* compartment or pore of length L (in the latter case it is ≈ the width of the pore ionic blockade or transverse current pulse) and with applied potential difference of V. From the Appendix

$$E(T) = (L^2/D\alpha)[1 - (1/\alpha)\,(1 - \exp(-\alpha))] \qquad (3)$$

and

$$\sigma^2(T) = (L^2/D\alpha^2)^2\,(2\alpha+4\alpha\exp(-\alpha)-5+4\exp(-\alpha)+\exp(-2\alpha)) \qquad (4)$$

where

$$v_z = \mu V/L; \qquad \alpha = v_z L/D = \mu V/D \qquad (5)$$

Here v_z is the drift velocity due to the electrophoretic force experienced by a charged particle in the z direction; it can be 0, negative, or positive. For $v_z = 0$, these two statistics are

$$E_0(T) = L^2/2D; \quad \sigma_0^2(T) = (1/6)\,(L^4/D^2) \qquad (6)$$

Details are given in Section A-1 of the Appendix, derivations may be found elsewhere.[11,12]

Let $T_{detector}$ = time resolution of the detector circuit (= ~1 μs with CMOS circuits[22]). The following are necessary conditions for correct sequencing:

C1: a) At most one cleaved fragment may occupy DNP (*Method 1*) or MNP (*Method 2*) at any time;
b) At most one cleaved residue may occupy DNP (*Method 2*).

C2: a) Cleaved fragments (*Method 1*) or residues (*Method 2*) must arrive at DNP in sequence order;
b) Cleaved fragments must arrive at MNP in sequence order (*Method 2*).

C3: a) A residue translocating through DNP must have a pulse width > $T_{detector}$ (*Method 2*);
b) A fragment with L_f residues must have a pulse width in DNP > $L_f T_{detector}$ (*Method 1*).

These conditions are influenced by the following factors:

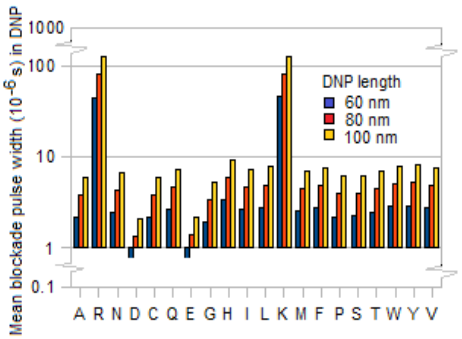
- The pore ionic blockade or transverse current pulse width, which is effectively the fragment or residue's residence time in DNP. It is approximated by the mean translocation time through DNP in both methods.
- The charge carried by a peptide fragment (and hence its mobility μ). As it depends on the constituent amino acids it has a wide range of values, which directly affects the translocation time (see Section A-2 in the Appendix for the relevant equations). Thus fragments with high negative charge have very high speeds of translocation which may result in misses ('deletes'), while those with high positive charge are 'lost' to diffusion because they are too slow. Figure A-1 in the Appendix shows the frequency distribution of all 20^7 peptides of length $L_F = 7$ as a function of μ or μ/D at pH = 7 (physiological pH), where D is the diffusion constant of the fragment. (Note the multimodal shape and slight negative skew.) These distributions are used in the Appendix to estimate the percentage of misses (deletes) and losses.

C1 and *C2* can be satisfied by requiring the enzymes to cleave at a given minimum rate. Enyzme reactions being stochastic processes, reaction rates are random variables with a distribution of values. The minimum rates required are estimated using standard statistical methods. *C3* can be satisfied through the use of a sufficiently thick pore. Thus the pore ionic blockade or transverse modulated current pulse width is proportional to the square of pore length (Equations 3 through 6), so a thicker pore can significantly increase translocation times and thus lower the required bandwidth (or equivalently increase the resolution time needed to sense the pulse). This is contrary to the usual practice of using thinner pores to achieve better discrimination,[6] but is appropriate here because residues do not have to be identified, they only have to be counted. (A side benefit of this is that thick synthetic pores are usually easier to fabricate than thin ones.[23]) See Discussion in Section 4.

With suitable values for the pore length, applied voltage, and peptide length all three conditions can be satisfied with a detector time resolution of ~1 μs. This is shown below by computing pulse widths and required enzyme reaction times. Only the results are given here, details may be found in the Appendix.

3.1 Computational results

The following parameter values are assumed: V_{05} = ~115 mV (*Method 1*); V_{07} = ~180 mV (*Method 2*); detector resolution = 1 μs; pore (DNP, MNP) conductance = ~1 nS; pH = 7.0; *trans*1/*cis*2 height = *trans*2/*cis*3 height = 0.5 μm, UNP length = MNP length = 10 nm. V_{05} divides as $V_{01} = V_{23} = V_{45} \approx 1.6$ mV, $V_{12} \approx 10$ mV, and $V_{34} \approx 100$ mV. V_{07} divides as $V_{01} = V_{23} = V_{45} = V_{67} \approx 1.5$ mV, $V_{12} = V_{34} \approx 15$ mV, and $V_{56} = 140$ mV. Let $T_{exo\text{-}min}$, $T_{endo\text{-}min\text{-}2}$, and $T_{endo\text{-}min\text{-}1}$ be the minimum reaction time intervals for the exopeptidase in *Method 2*, the endopeptidase in *Method 2*, and the endopeptidase in *Method 1* respectively. Translocation time distributions are assumed to have 6σ support, where σ is the standard deviation.



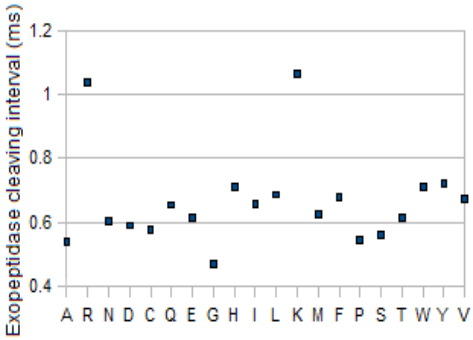


Figure 4 (left). Mean blockade pulse width (μs) in DNP of different lengths for the 20 individual amino acids in *Method 2*. *trans2* height = 0.5 μm, V_{56} = 140 mV, V_{45} = 1.2 mV, pH = 7.

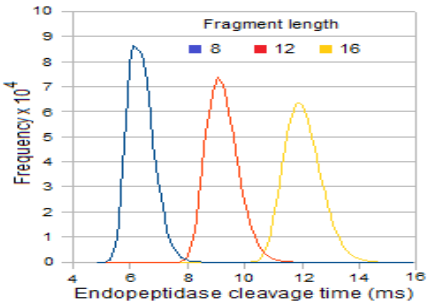
Figure 5 (right). Distribution of $T_{exo\text{-}min}$ (ms) for each amino acid type in *Method 2*. DNP height = 80 nm, *trans*2 height = 0.5 μm, V_{56} = 140 mV, V_{45} = 1.2 mV, pH = 7.

Method 2: Using data from Table A-3 in the Appendix for DNP length = 80 nm, the minimum mean blockade pulse width in DNP is given by the fastest amino acid (Asp) and is ~1.33 μs > $T_{detector}$ = 1 μs. $T_{exo\text{-}min}$ is largely determined by the slowest (Lys) and is ~1 ms. More generally Figure 4 shows the mean blockade pulse widths due to single residues in DNP for all 20 residue types for three different lengths of DNP, while Figure 5 shows $T_{exo\text{-}min}$ vs residue type for DNP length = 80

nm. Figure 6 shows the frequency distribution of $T_{endo\text{-}min\text{-}2}$ for different fragment lengths (L_F), each based on 10^6 randomly generated peptide sequences. In each case ~95% of the sequences have minimum enzyme cleavage intervals < L_F ms.

Method 1: Figure 7 shows the distribution of fragment pulse widths for three different fragment lengths. Figure 8 shows $T_{endo\text{-}min\text{-}1}$ for 10^6 random samples of length 12 in *Method 1*. For the vast majority of sequences $T_{endo\text{-}min\text{-}1}$ is < 3 ms. The curve is to the left of the corresponding curve in *Method 2* (Figure 6, red) because the endopeptidase reaction times in the latter include the delay due to the cleaving of residues in a fragment by the exopeptidase (although this is not strictly necessary because the pulses are only counted so they can arrive in any order). The distribution of pulse widths > 12 µs due to fragments of length = 12 vs the endopeptidase reaction time is shown in Figure 9. For nearly 80% of the sequences (with pulses in which L_F transitions can be counted) $T_{endo\text{-}min\text{-}1}$ < 1 ms. In comparison the percentage of pulses that may not be counted correctly is relatively small at ~17%. The curves in Figures 6 and 8 are similar in shape and range to reaction rate graphs for the enzyme Exonuclease I.[24]

Comparing the two methods. Method 1 has a more compact physical structure and uses only one enzyme, but the need to recognize transitions in a blockade pulse due to a fragment reduces the maximum length that can be determined accurately. The ionic current is also lower. *Method 2* can use a shorter (that is, thinner) DNP and a higher potential difference (leading to a higher ionic current). (This is not as serious a problem with transverse currents, which are on the order of nA,[23] compared with at most 100s of pA with ionic currents.) However, as noted earlier, the reaction time required of the endopeptidase is significantly larger; also the endopeptidase and exopeptidase need to cleave at a sufficiently low rate and in synchrony. Notice that in *Method 1* even if the exopeptidase is inefficient and does not cleave after every single residue, the number of residues would be counted correctly if DNP detects transitions between residues in a pulse due to a fragment with more than one residue.



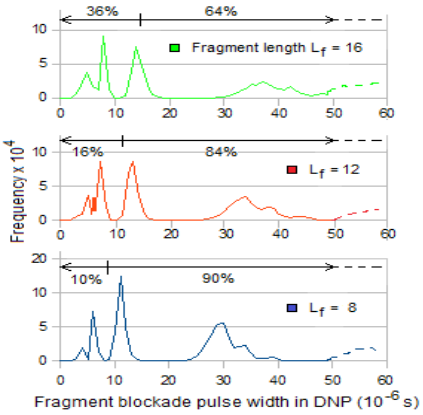
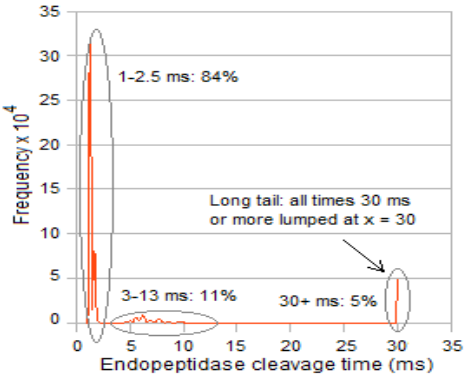


Figure 6 (left) . Frequency distribution of $T_{endo\text{-}min\text{-}2}$ (ms) for different fragment lengths L_F in *Method 2*. MNP height = 10 nm, trans1 height = 0.5 µm, V_{23} = 1.2 mV, V_{34} = 20 mV, pH = 7.

Figure 7 (right). Frequency distribution of fragment pulse widths (µs) for different fragment lengths in *Method 1*. DNP height = 150 nm, *trans*1 height = 0.5 µm, V_{34} = 100 mV, V_{23} = 1.2 mV, pH = 7.



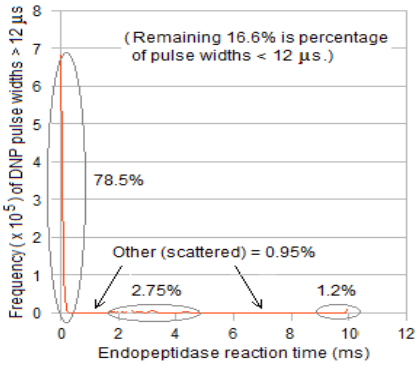


Figure 8 (left). Frequency distribution of $T_{endo\text{-}min\text{-}1}$ (ms) for fragment length = 12 in *Method 1*. DNP height = 150 nm, *trans*1 height = 0.5 µm, V_{34} = 100 mV, V_{23} = 1.2 mV, pH = 7.

Figure 9 (right). Distribution of pulse widths > 12 µs for fragment length = 12 vs $T_{endo\text{-}min\text{-}1}$ (ms) in *Method 1*. DNP height = 150 nm, *trans*1 height = 0.5 µm, V_{34} = 100 mV, V_{23} = 1.2 mV, pH = 7.

4 Discussion

Some relevant implementation issues are considered next.

1) *Counting pulses or transitions in a pulse*. On the face of it counting transitions in a pulse due to residues in a fragment would appear to be easier with the following methods: a) using single-atom thick graphene[25] or molybdenum disulphide (MoS_2) sheets,[26] both of which make counting of transitions easier; b) detecting level crossings in a transverse electronic[26] or tunneling[27,28] current pulse across graphene or silicon gaps; and c) using a narrow biological nanopore like MspA, which has a constriction in its short stem[29] that may aid in recognizing the transitions. However all of these methods would require bandwidths in the tens of MHz if directly used in the approach described here. To bring the bandwidth down to 1-2 MHz (corresponding to a pulse width resolution of ~1 μs), thick pores may be considered, as discussed in Section 3. With silicon compounds like Si_3N_4 thick pores of 50-100 nm are actually easier to manufacture than thin pores.[23,30] With graphene, hourglass-shaped pores may be fabricated from graphite (which is a stack of graphene layers[31]) but stability may be an issue because of graphite's flakiness. Biological pores like AHL or MspA can also be stacked, for example a stack of 10 AHL pores can provide a pore about 60-80 nm thick.

2) *Location of peptidases*. The cleaving action of an enzyme (endopeptidase or exopeptidase) requires it to be in the path of the peptide or fragment emerging from the respective pore (UNP or MNP) on the *trans* side. This can be ensured by covalently attaching the enzyme to the *trans* side of the pore membrane. Such covalent attachment has been discussed for DNA sequencing in two different approaches: exosequencing of mononucleotides[32] and sequencing by synthesis using heavy tags attached to the bases.[33] In both approaches an exonuclease or polymerase is attached to the *cis* side of the pore membrane. This could result in significant errors due to cleaved bases or tags being lost to diffusion in the *cis* chamber (deletions) or entering the pore out of order (delete-and-insert).[34] In the present approach the peptidases are located on the *trans* side so deletions cannot occur. Out-of-order arrivals at the sensing pore (fragments at DNP in *Method 1*; fragments at MNP and residues at DNP in *Method 2*) are precluded as long as the necessary conditions given in Section 3 are satisfied.

3) *Solution pH*. Solution pH plays an important role for two reasons: a) the charge carried by a fragment, which is highly variable and not known in advance, is a function of pH; compare with DNA, where all nucleotide types have approximately the same electron charge of -q with a small variability due to pH; b) its effect on enzyme reaction rates. The choice of pH is a tradeoff between enzyme efficiency and being able to control translocation speeds; this may be determined by experiment.

4) *Fabrication*. A recent review of nanopore sequencing includes notes on fabrication techniques.[30] Recently a tandem-pore-like structure was used to trap and analyze DNA,[35] with the *trans*1/*cis*2 chamber functioning like a test-tube. In contrast with conventional nanopore sequencing methods, where the aim is to fabricate thin pores (~1 nm) that are usually synthetic (such as, for example, Si_3N_4), as noted earlier the thick (80-100 nm) pores required in the proposed scheme may be more easily fabricated.

5) *Other*. See Appendix.

Supplementary information. Two data files containing the number of identifying sequences for each protein in the human proteome, one for each of two sets of cleaving options, are available.

Email: sampath_2068@yahoo.com

References

[1] P. C. Ng and E. F. Kirkness, "Whole genome sequencing," *Methods Mol. Biol.*, 2010, **628**, 215-226.
[2] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry*, 7th edn., New York, W H Freeman, 2012.
[3] H. Steen and M. Mann. "The ABC'S (and XYZ's) of peptide sequencing." *Nature Reviews*, 2004, **5**, 699-711.
[4] J. Quick, A. Quinlan, and N. Loman. "A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer." *Gigascience*, 2014, **3**, 22-27.
[5] M. Yang, T.-Y. Kim, H.-C. Hwang, S.-K. Yi, D.-H. Kim, "Development of a palm portable mass spectrometer," *J. Am. Soc. Mass Spectrom*. **2008**, *19*, 1442–1448.
[6] M. Wanunu, "Nanopores: a journey towards DNA sequencing." *Phys Life Rev*, 2012, **9**, 125–158.
[7] W. Timp, A. M. Nice, E. M. Nelson, V. Kurz, K. Mckelvey, and G. Timp. "Think small: nanopores for sensing and synthesis." *IEEE Access*, 2014, **2**, 1396-1408.
[8] A. Oukhaled, L. Bacri, M. Pastoriza-Gallego, J-M. Betton, and J. Pelta, "Sensing proteins through nanopores: fundamental to applications," *ACS Chem. Biol*., 2012, **7**, 1935-1949.
[9] D. Wu, S. Bi, L. Zhang, and J. Yang. "Single-molecule study of proteins by biological nanopore sensors." *Sensor*, 2014, **14**, 18211-18222.
[10] D. Rotem, L. Jayasinghe, M. Salichou, and H. Bayley, "Protein detection by nanopores equipped with aptamers", *J. Amer. Chem. Soc*., 2012, **134**, 2781-2787.
[11] G. Sampath, "A tandem cell for nanopore-based DNA sequencing with exonuclease," *RSC Adv.*, 2015, **5**, 167-171.
[12] G. Sampath, "Amino acid discrimination in a nanopore and the feasibility of sequencing peptides with a tandem cell and exopeptidase," *RSC Adv.*, 2015, **5**, 30694-30700.
[13] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *PNAS*, 1977, **74**, 560-564.
[14] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *PNAS*, 1977, **74**, 5463–5467.
[15] J. Swaminathan, A. A. Boulgakov, E. M. Marcotte, "A theoretical justification for single molecule peptide sequencing," *PLoS Comput. Biol*., 2015, **11**, e1004080.
[16] http://web.expasy.org/peptide cutter/. Accessed July 11, 2015.

[17] K. Breddam, "Serine carboxypeptidases: a review," *Carlsberg Res. Commun*., 1986, 51, 83-128.
[18] K. Breddam and M. Ottesen, "Determination of c-terminal sequences by digestion with serine carboxypeptidases: the influence of enzyme specificity," *Carlsberg Res. Commun*., 1987, **52**, 55-63.
[19] A. Taylor, "Aminopeptidases: structure and function," *FASEB J*., 1993, 7, 290-298.
[20] http://www.uniprot.org/uniprot. Accessed July 4, 2015.
[21] D. Gusfield. *String Algorithms*. Cambridge University Press, Cambridge (UK), 1997.
[22] J. K. Rosenstein, M. Wanunu, C. A. Merchant, M, Drndic, and K. L. Shepard, "Integrated nanopore sensing platform with sub-microsecond temporal resolution," *Nature Methods*, 2012, **9**, 487–492.
[23] A. Balan, B. Machielse, D. Niedzwiecki, J. Lin, P. Ong, R. Engelke, K. L. Shepard, and M. Drndic´, "Improving signal-to-noise performance for DNA translocation in solid-state nanopores at MHz bandwidths," *Nano Lett*., **2014**, *14*, 7215-7220.
[24] J. H. Werner, H. Cai, R. A. Keller, P. M. Goodwin, "Exonuclease I hydrolyzes DNA with a distribution of rates," *Biophys. J*., **2005**, *88*, 1403–1412.
[25] M. Drndic, "Sequencing with graphene pores," *Nat. Nanotech*., **2014**, *9*, 743.
[26] A. B. Farimani, K. Min, N. R. Aluru, "DNA base detection using a single-layer MoS_2", *ACS Nano*, **2014**, *8*, 7914-7922.
[27] M. Tsutsui, M. Taniguchi, K. Yokota, T. Kawai, "Identifying single nucleotides by tunnelling current," *Nat. Nano*, **2010**, *5*, 286–90.
[28] Y. Zhao, B. Ashcroft, P. Zhang, H. Liu, S. Sen, W. Song, J. Im, B. Gyarfas, S. Manna, S. Biswas, C. Borges, and S. Lindsay, "Single-molecule spectroscopy of amino acids and peptides by recognition tunneling," *Nature Nanotech*., 2014, **9**, 466–473.
[29] T. Z. Butler, M Pavlenok, I. M. Derrington, M Niederweis, J. H. Gundlach, "Single-molecule DNA detection with an engineered MspA protein nanopore," *PNAS*, **2008**, *105*, 20647–20652.
[30] Y. Wang, Q. Yang, Z. Wang, "The evolution of nanopore sequencing," *Front. Genet*., **2015**, *5*, 449.
[31] A. van der Zande, "The structure and mechanics of atomically-thin graphene membranes," PhD thesis, **2011**, Cornell University.
[32] J. Clarke, H-C. Wu, L Jayasinghe, A. Patel, S. Reid and H. Bayley, "Continuous base identification for single-molecule nanopore DNA sequencing," *Nature Nanotech*., **2009**, *4*, 265-270.
[33] S. Kumar, C. Tao, M. Chien, B. Hellner, A. Balijepalli, J. W. F. Robertson, Z. Li, J. J. Russo, J. E. Reiner, J. J. Kasianowicz, and J. Ju, "PEG-labeled nucleotides and nanopore detection for single molecule DNA sequencing by synthesis," *Scientific Reports*, **2012**, DOI:10.1038/srep00684.
[34] J. E. Reiner, A. Balijepalli, J. F. Robertson, D. L. Burden, B. S. Drown, and J. J. Kasianowicz, "The effects of diffusion on an exonuclease/nanopore-based DNA sequencing engine," *J. Chem Phys*. **2012**, *137*, 214903.
[35] X. Liu, M. M. Skanata, D. Stein, "Entropic cages for trapping DNA near a nanopore," *Nat. Comms*. **2015**, *6*, doi:10.1038/ncomms7222.

Appendix

A-1 Translocation statistics of tandem cell

Following [11,12], the mean E(T) and variance $\sigma^2(T)$ of the translocation time T over a channel of length L that is reflective at the top and absorptive at the bottom with applied potential difference of V are given by

$$E(T) = (L^2/D\alpha)[1 - (1/\alpha)\,(1 - \exp(-\alpha))] \qquad \text{(A-1)}$$

and

$$\sigma^2(T) = (L^2/D\alpha^2)^2\,(2\alpha + 4\alpha\exp(-\alpha) - 5 + 4\exp(-\alpha) + \exp(-2\alpha)) \qquad \text{(A-2)}$$

with $$v_z = \mu V/L; \qquad \alpha = v_z L/D = \mu V/D \qquad \text{(A-3)}$$

Here v_z is the drift velocity due to the electrophoretic force experienced by a charged particle in the z direction, which can be 0, negative, or positive. For $v_z = 0$, these two statistics are

$$E_0(T) = L^2/2D; \qquad \sigma_0^2(T) = (1/6)\,(L^4/D^2) \qquad \text{(A-4)}$$

If each section in the double tandem cell is considered independently these formulas can be applied to all the relevant sections: *trans*1/*cis*2 ($T = T_{trans1/cis2}$; $L = L_{23}$), MNP ($T = T_{MNP}$; $L = L_{34}$), *trans2*/*cis3* ($T = T_{trans2/cis3}$; $L = L_{45}$), DNP ($T = T_{DNP}$; $L = L_{56}$), and *trans*3 ($T = T_{trans3}$; $L = L_{67}$). For an analysis of behavior at the interface between two sections see [11,12].

A-2 Dependence of particle translocation on solution pH, charge, diffusion constant, and mobility

Equations A-1 through A-4 involve a number of physical-chemical properties of amino acids: electrical charge (itself dependent on solution pH) [36], hydrodynamic radius, diffusion constant, and mobility. The following paragraphs provide a quantitative description of this dependence and allow calculation of fragment properties as they apply to peptide sequencing in a tandem cell with endopeptidase. In particular this information is used in the next section to derive a required condition for effective sequencing.

Table A-1

Peptide end or amino acid	Amino end	Carboxy end	R	D	C	E	H	K	Y
kA value	2.34	9.69	12.48	3.86	8.33	4.25	6.0	10.53	10.07

1) The electrical charge carried by a peptide (fragment) P_x can be calculated with the Henderson-Hasselbach equation. Let the set of amino acids be **AA** = [A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V] where AA[i] is the i-th amino acid, $1 \le i \le 20$. Let the pH value of the solution (electrolyte) be p, kC = kA value of the carboxy end = 9.69, kN = kA value of the amino end = 2.34, N_X the number of times residue X occurs in the peptide (X = R, H, K), N_Z the number of times residue Z occurs (Z = D, C, E, Y), and kX and kZ the kA values of X and Z respectively. kA values are given by Table A-1. The charge multiplier C_{Px} on the peptide is given by

$$C_{Px} = 10^{kC} / (10^{p} + 10^{kC}) - 10^{p} / (10^{p} + 10^{kN}) + \sum_X 10^{kX} / (10^{p} + 10^{kX}) - \sum_Z 10^{p} / (10^{p} + 10^{kZ}) \qquad \text{(A-5)}$$

where the summations are over the N_X and N_Z occurrences of X and Z respectively in P_x.

2) The hydrodynamic radius R_{Px} of peptide $P_x = X_1 X_2 \ldots X_N$ is obtained recursively as follows:

$$\begin{aligned} R_{X1 \ldots Xk} &= R_{X1 \ldots X\,k-1} (1 + 3 (V_{Xk} - \delta v/2) / 4\pi (R_{X1 \ldots X\,k-1})^3))^{1/3}, \quad k > 1 \\ &= R_{X1}, \quad k = 1 \end{aligned} \qquad \text{(A-6)}$$

where V_{Xk} and and δv are the van der Waals volumes of X_k and a single molecule of water. Hydrodynamic radii of individual amino acids are given in [37] and van der Waals volumes in [38] (both sets of values are reproduced in the Supplement to [12]). This formula holds for small peptides (up to ~20 residues).

3) The diffusion constant and mobility of P_x are given by

$$D_{Px} = k_B T_R / 6\pi\eta R_{Px} \qquad \mu_{Px} = C_{Px}\, q / 6\pi\eta R_{Px} \qquad \text{(A-7)}$$

Here k_B is the Boltzmann constant (1.3806×10^{-23} J/K), T_R is the room temperature (298° K), η is the solvent viscosity (0.001 Pa.s), q is the electron charge (1.619×10^{-19} coulomb), and C_{Px} is a multiplier.

Figure 1 shows the distribution of the number of peptides of length 7 vs mobility μ and μ/D (= α with V set to 1) over all 20^7 of them.

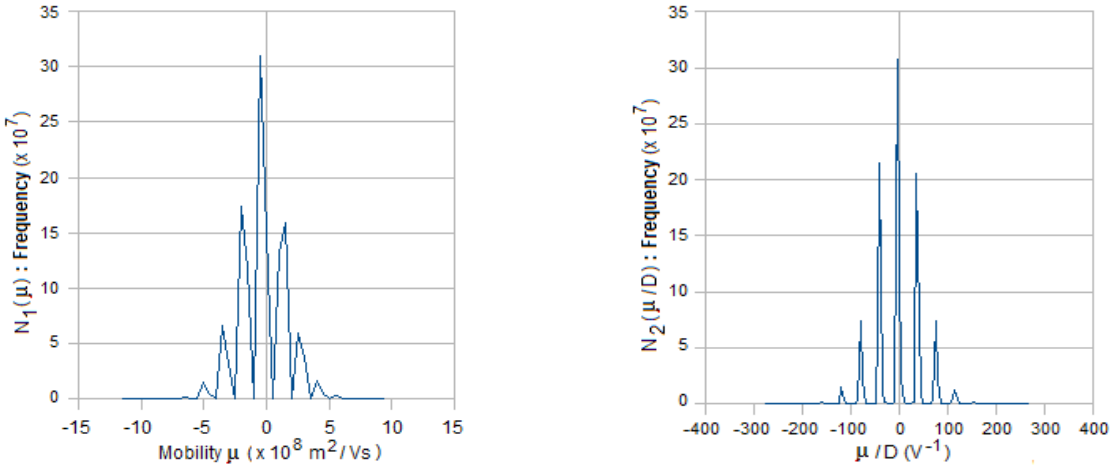


Figure A-1. Distribution of number of peptides of length 7 vs mobility μ and the ratio of mobility to diffusion coefficient (μ/D).

A-3 Calculating the percentage of misses (deletes) due to fast fragments and losses due to slow fragments

For fragments that carry a high negative or positive charge the mean translocation time in Equations A-1 and A-4 can be approximated by

$$E(T) \approx L^2 \mu V, \quad \alpha >> 0 \qquad \text{(A-8a)}$$

$$\approx L^2 \exp(-\alpha) / D\,\alpha^2, \quad \alpha << 0 \qquad \text{(A-8b)}$$

These formulas can be used to estimate the percentage of misses due to fast translocating fragments and slowly moving fragments through DNP in Method 1.

To estimate the former, for a given pore length L, voltage V across the pore, and blockade pulse width approximated by E(T), μ is written as

$$\mu = E(T) / L^2 V \qquad \text{(A-9)}$$

The percentage of misses is given by

$$\% \text{ misses} = 100(1/20^7)\int_{-\infty}^{\mu} N_1(\mu)\, d\mu \qquad \text{(A-10)}$$

The integral in Equation A-10 is the cumulative frequency for $N_1(\mu)$ corresponding to the μ calculated from Equation A-9. The results are shown in Table A-2 for V = 100 mV, L = 100, 150, 200, and 250 nm, and two pulse widths: E(T) = 7 μs and 10 μs.

To estimate the percentage of losses rewrite Equation A-8b as

$$E(T) = L^2 \exp(-\mu V/D) / D(-\mu V/D)^2 \qquad \text{(A-11)}$$

This is an implicit function of two parameters, μ/D and D. To solve for μ/D for a given E(T), L, and V, it is approximated by

$$E(T) \approx L^2 \exp(-\mu V/D) / D_{avg}(-\mu V/D)^2 \qquad \text{(A-12)}$$

where D_{avg} is the average diffusion coefficient of all 20^7 peptides of length 7. This is a nonlinear equation in μ/D; the desired root on the real line can be found using standard methods. For a given value of V, the percentage of losses is given by

$$\% \text{ loss} = 100(1/20^7)\int_{\mu/D}^{\infty} N_2(\mu/D)\, d(\mu/D) \qquad \text{(A-13)}$$

The results are shown in Table A-2 for V = 100 mV and L = 100, 150, 200, and 250 nm, and E(T) = 1 s.

Table A-2

L (nm)	Pulse width = 7 μs			Pulse width = 10 μs			Pulse width = 1 s		
	μ (m^2/Vs)	N_1 (μ)	Percentage deletions	μ (m^2/Vs)	N_1 (μ)	Percentage deletions	μ/D (V^{-1})	N_2 (μ/D)	Percentage losses
100	-1.43	4.1×10^8	32.09	-1	7.2×10^8	56.27	1.6×10^2	2.06×10^5	0.02
150	-3.22	1.19×10^8	9.3	-2.25	2.93×10^8	22.89	1.52×10^2	2.52×10^8	0.19
200	-5.72	2.97×10^6	0.23	-4	2.2×10^7	1.72	1.45×10^2	2.52×10^8	0.19
250	-8.94	7.4×10^3	0	-6.25	2.5×10^6	0.2	1.4×10^2	2.52×10^8	0.19

Figures A-2 and A-3 show similar distributions for peptide lengths 12 and 16.

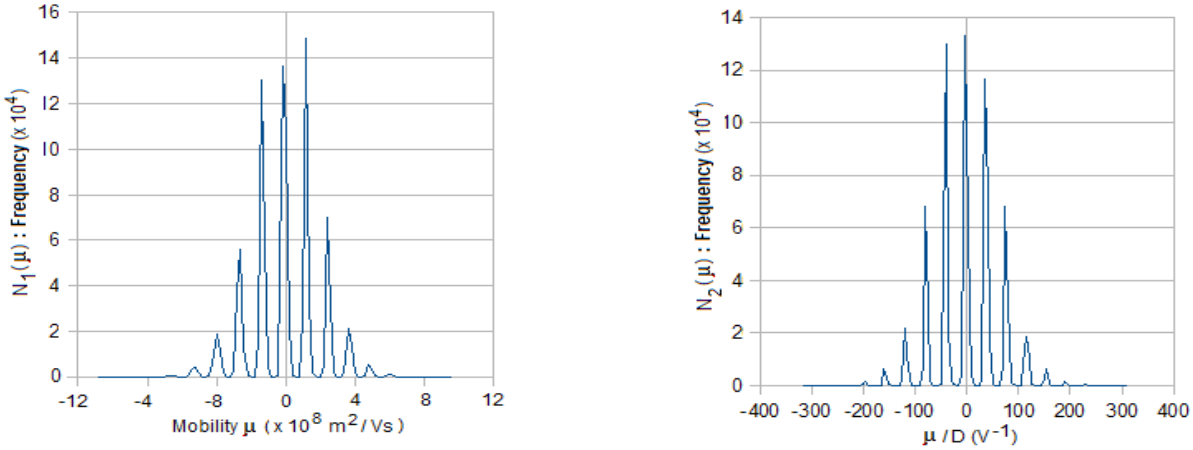


Figure A-2. Distribution of number of peptides vs mobility μ and the ratio of mobility to diffusion coefficient (μ/D). Numbers based on 10^6 randomly generated peptide strings of length 12.

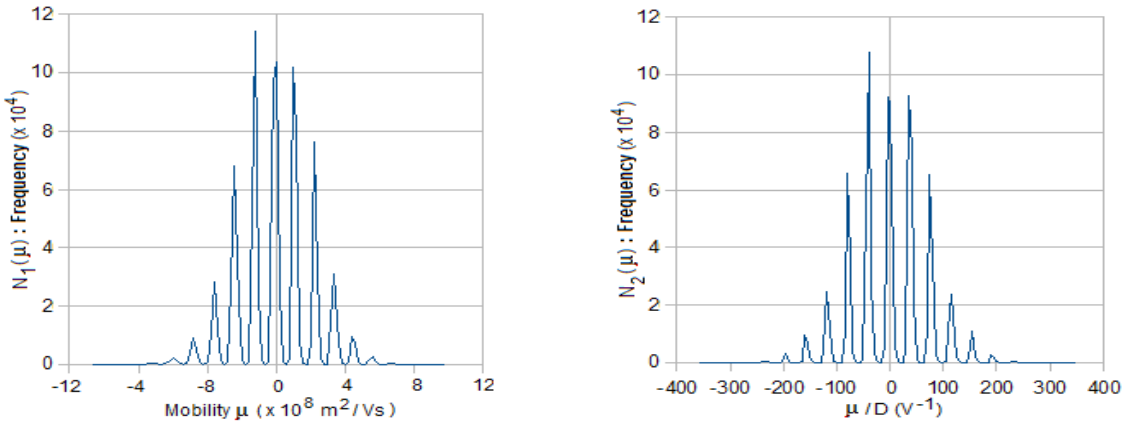


Figure A-3. Distribution of number of peptides vs mobility μ and the ratio of mobility to diffusion coefficient (μ/D). Numbers based on 10^6 randomly generated peptide strings of length 16.

A-4 Translocation statistics of single residues

The mean and standard deviation of the time taken by a single residue through *trans*2/*cis*3 and DNP (Method 2) are shown in Table A-3 as a function of pH.

Table A-3

Method 2: Translocation time of single residues through *trans*2/*cis*3 (10^{-3} s) and DNP (10^{-6} s)

Amino acid	pH=3				pH=7				pH=11			
	*trans*2/*cis*3		DNP		*trans*2/*cis*3		DNP		*trans*2/*cis*3		DNP	
	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
A	0.1528	0.1248	5.5557	4.8187	0.1523	0.1244	3.8857	3.1704	0.1501	0.1222	1.2137	0.6792
R	0.2100	0.1721	156.44	155.54	0.2094	0.1715	79.93	78.98	0.2062	0.1677	5.4214	4.4501
N	0.1711	0.1398	6.2241	5.3984	0.1707	0.1394	4.3532	3.5518	0.1682	0.1369	1.3597	0.7609
D	0.1731	0.1414	4.9358	4.1143	0.1703	0.1386	1.3264	0.7297	0.1678	0.1362	0.7536	0.3153
C	0.1643	0.1342	5.9734	5.1809	0.1637	0.1336	3.8589	3.0969	0.1589	0.1290	0.7143	0.2990
Q	0.1855	0.1516	6.7462	5.8513	0.1850	0.1510	4.7184	3.8497	0.1823	0.1484	1.4738	0.8248
E	0.1802	0.1472	5.8700	5.0063	0.1771	0.1441	1.3803	0.7596	0.1745	0.1416	0.7835	0.3278
G	0.1332	0.1089	4.8456	4.2028	0.1329	0.1085	3.3891	2.7651	0.1309	0.1066	1.0586	0.5924
H	0.2036	0.1668	151.08	150.21	0.2002	0.1635	6.0541	5.0998	0.1969	0.1603	1.5924	0.8912
I	0.1861	0.1520	6.7671	5.8694	0.1856	0.1515	4.7330	3.8617	0.1828	0.1488	1.4783	0.8273
L	0.1947	0.1591	7.0804	6.1411	0.1942	0.1585	4.9521	4.0404	0.1913	0.1557	1.5468	0.8656
K	0.2153	0.1764	160.35	159.42	0.2147	0.1758	81.84	80.87	0.2090	0.1703	2.1084	1.2987
M	0.1769	0.1445	6.4329	5.5795	0.1764	0.1440	4.4993	3.6710	0.1738	0.1415	1.4053	0.7865
F	0.1924	0.1572	6.9969	6.0686	0.1919	0.1567	4.8937	3.9928	0.1890	0.1539	1.5285	0.8554
P	0.1539	0.1257	5.5975	4.8549	0.1535	0.1253	3.9150	3.1942	0.1512	0.1231	1.2228	0.6843
S	0.1585	0.1295	5.7646	4.9998	0.1581	0.1291	4.0318	3.2896	0.1557	0.1268	1.2593	0.7048
T	0.1746	0.1426	6.3494	5.5071	0.1741	0.1422	4.4409	3.6233	0.1716	0.1397	1.3871	0.7763
W	0.2010	0.1642	7.3102	6.3404	0.2005	0.1637	5.1128	4.1715	0.1975	0.1608	1.5970	0.8937
Y	0.2050	0.1675	7.4564	6.4672	0.2045	0.1669	5.2070	4.2471	0.1987	0.1613	0.9359	0.4014
V	0.1907	0.1558	6.9342	6.0143	0.1901	0.1553	4.8499	3.9570	0.1874	0.1525	1.5148	0.8477

trans2/*cis3*: height = 0.5 μm radius = 0.5 μm V_{23} = 1.2 mV DNP: height = 80 nm radius = 2 nm V_{34} = 140 mV

A-5 Necessary conditions for effective sequencing

The material between << and >> is repeated from the main text.

<<

Let $T_{detector}$ be the time resolution of the detector circuit (= ~1 μs with CMOS circuits[22]). The following are necessary conditions for

effective sequencing:

C1: a) At most one cleaved fragment may occupy DNP (*Method 1*) or MNP (*Method 2*) at any time;
b) At most one cleaved residue may occupy DNP (*Method 2*).

C2: a) Cleaved fragments (*Method 1*) or residues (*Method 2*) must arrive at DNP in sequence order;
b) Cleaved fragments must arrive at MNP in sequence order (*Method 2*).

C3: a) A residue translocating through DNP must have a pulse width > $T_{detector}$ (*Method 2*);
b) A fragment with L_f residues must have a pulse width in DNP > $L_f T_{detector}$ (*Method 1*).

>>

It is now shown that the conditions applicable to each of the two methods are satisfied by a large majority (~80% in most cases) of peptide sequences of a given length for a set of typical parameter values. In the following translocation time distributions are assumed to have 6σ support (σ = standard deviation).

Method 2. *Conditions 1a*, *1b*, *2a*, *2b*, and *3* have to be satisfied. From Table A-3, with pH = 7.0, DNP height = 80 nm, and V_{56} = 140 mV, the fastest amino acid is Asp (D) with a translocation time of ~1.33 μs > 1 μs. This satisfies *Condition 3*.
Let X_1 and X_2 be two residues cleaved in succession by the exopeptidase. *Conditions 1a* , *1b*, and *2a* are satisfied if

$$E(T_{trans2/cis3\text{-}X1}) + 3\sigma_{trans2/cis2\text{-}X!} + E(T_{DNP\text{-}X1}) + 3\sigma_{DNP\text{-}X1} < T_{exo\text{-}min} + \max\,(0,\, E(T_{trans2/cis3\text{-}X2}) - 3\sigma_{trans2/cis3\text{-}X2}) \qquad \text{(A-14)}$$

From columns 6 and 7 in the same table the second term in the inequality on the right is 0, leading to

$$T_{exo\text{-}min} > \max_X \{\, E(T_{trans1/cis2\text{-}X}) + 3\sigma_{trans1/cis2\text{-}X} + E(T_{DNP\text{-}X}) + 3\sigma_{DNP\text{-}X} \,\} \qquad \text{(A-15)}$$

over all X. The maximum occurs for X = K (Lys), with $E(T_{trans2/cis3\text{-}X}) = 0.21\times10^{-3}$, $\sigma_{trans2/cis3\text{-}X} = 0.18\times10^{-3}$, $E(T_{DNP\text{-}X}) = 82\times10^{-6}$, and $\sigma_{DNP\text{-}X} = 81\times10^{-6}$, leading to

$$T_{exo\text{-}min} \approx 1 \text{ ms} \qquad \text{(A-16)}$$

More generally the rate can be calculated for each residue type in a similar way. Figure 5 in the main text shows the required minimum cleaving interval with DNP height = 80 nm, V_{56} = 140 mV, and V_{45} = 1.2 mV.

A peptide that has threaded through UNP encounters the endopeptidase in *trans*1/*cis*2 and is cleaved into fragments. The latter translocate through *trans*1/*cis*2 and thread through MNP to be cleaved by the exopeptidase on the downstream side. Consider two successive fragments F_1 and F_2. Let L_{F1} be the length of a fragment F_1. The delay due to cleaving of F_1 into single residues by the exopeptidase is $L_{f1} T_{exo\text{-}min\text{-}2}$. *Conditions 1a* and *2b* will be satisfied if

$$E(T_{trans1/cis2\text{-}F1}) + 3\sigma_{trans1/cis2\text{-}F1} + E(T_{MNP\text{-}F1}) + 3\sigma_{MNP\text{-}F1} + \textstyle\sum_{F1} T_{exo\text{-}min\text{-}X} < T_{endo\text{-}min\text{-}2} + \max\,(0,\, E(T_{trans1/cis2\text{-}F2}) - 3\sigma_{trans1/cis2\text{-}F2}) \qquad \text{(A-17)}$$

where $T_{exo\text{-}min\text{-}X}$ is the cleavage time for residue X and the summation is over all L_{F1} residues in F_1. In the second term on the right side of the inequality, $\sigma_{trans1/cis2\text{-}F2} \approx E(T_{trans1/cis2\text{-}F2})$, so that $\max\,(0, E(T_{trans1/cis2\text{-}F2}) - 3\sigma_{trans1/cis2\text{-}F2}) = 0$; this leads to

$$T_{endo\text{-}min\text{-}2} = E(T_{trans1/cis2\text{-}F1}) + 3\sigma_{trans1/cis2\text{-}F1} + E(T_{MNP\text{-}F1}) + 3\sigma_{MNP\text{-}F1} + \textstyle\sum_{F1} T_{exo\text{-}min\text{-}X} \qquad \text{(A-18)}$$

Figure 6 in the main text shows the distribution of $T_{endo\text{-}min\text{-}2}$ with 10^6 random peptide sequences with residues in a sequence drawn from a uniform distribution for three different fragment lengths.

Method 1. The development is similar to that for *Method 2*. Thus *Conditions 1a*, *2a*, and *3* have to be satisfied. With two successive fragments F_1 and F_2 cleaved by the endopeptidase, *Conditions 1a* and *2a* require

$$E(T_{trans1/cis2\text{-}F1}) + 3\sigma_{trans1/cis2\text{-}F1} + E(T_{DNP\text{-}F1}) + 3\sigma_{DNP\text{-}F1} < T_{endo\text{-}min\text{-}1} + \max\,(0,\, E(T_{trans1/cis2\text{-}F2}) - 3\sigma_{trans1/cis2\text{-}F2}) \qquad \text{(A-19)}$$

As before the second term on the right is 0 because $\sigma_{trans1/cis2\text{-}F2} \approx E(T_{trans1/cis2\text{-}F2})$ which leads to

$$T_{endo\text{-}min\text{-}1} = E(T_{trans1/cis2\text{-}F1}) + 3\sigma_{trans1/cis2\text{-}F1} + E(T_{DNP\text{-}F1}) + 3\sigma_{DNP\text{-}F1} \qquad \text{(A-20)}$$

Figure 7 in the main paper shows the pulse width distribution for L_f = 8, 12, and 16 based on 10^6 random samples of peptide sequences with residues drawn from a uniform distribution. In each case the percentage of pulse widths < L_f μs (= L_f $T_{detector}$ with $T_{detector}$ = 1 μs) is also indicated; these correspond to deletes. Figure 8 shows the distribution of $T_{endo\text{-}min\text{-}1}$ for a fragment length of 12 for 10^6 random samples of length 12. The distribution of pulse widths due to fragments of length > 12 μs vs the endopeptidase reaction time is shown in Figure 9.

A-6 Peptidases and chemicals for cleaving and their specificities

Table A-4 is a summary of selected chemicals and peptidases for use in cleaving of the unknown protein or peptides generated from it at desired locations; it is adapted from [16]. The following notation is used for cleavage sites on a substrate [39]:

$$N - P_n - \ldots - P_4 - P_3 - P_2 - P_1 = \text{cut} = P_1' - P_2' - P_3' - P_4' - \ldots - P_m' - C$$

where – represents a peptide bond, N is the N-terminal end, and C is the C-terminal end.

Table A-4

Selected chemical agents for cleaving a protein after or before a specific amino acid (Stage 1)

Chemical	Cleavage point and target	Chemical	Cleavage point and target(s)
BNPS-Skatole	N – ... – Trp =cut= ... – C	Hydroxylamine	N – ... – Asn =cut= Gly – ... – C
Cyanogen bromide (CNBr)	N – ... – Met =cut= ... – C	Iodosobenzoic acid	N – ... – Trp =cut= ... – C
Formic acid	N – ... – Asp =cut= ... – C	NTCB +Ni	N – ... =cut= Cys – ... – C

Selected peptidases for cleaving a peptide after or before one or more specific amino acids (Stage 1 or Stage 2)
(Alternative target, usually less probable,is in parentheses.)

Peptidase	Cleavage point	Target X	Peptidase	Cleavage point	Target X
ArgC proteinase	N – ... – X =cut= ... – C	X = Arg	ArgC endopeptidase	N – ... – X =cut= ... – C	X = Arg (Lys)
AspN endopeptidase	N – ... =cut= X – ... – C	X = Asp (Glu)	LysC endopeptidase	N – ... – X =cut= ... – C	X = Lys (Asn)
Trypsin	N – ... X =cut= Z – ... – C	X = Arg or Lys Z = not Pro	LysC lysyl endopeptidase	N – ... – X =cut= ... – C	X = Lys
LysN peptidyl metalloendopeptidase	N – ... =cut= X – ... – C	X = Lys	Neutrophil elastase	N – ... – X =cut= ... – C	X = Val or Ala
Glutamyl endopeptidase	N – ... – X =cut= ... – C	X = Glu			

A-7 Additional notes

1) *Order of fragment entry into DNP*. A fragment can enter DNP amino-end first or carboxy-end first. However the order is not important as the information sought is the number of residues, not their identity or sequence.

2) *Order of entry of peptide into UNF*. The assembly algorithm described in Section 2 implicitly assumes that entry of a peptide into UNP in each of the cells is all of them either N-terminal first or C-terminal first. This is a reasonable assumption because of the charged X-header. However, there is a non-zero probability that the peptide may enter wrong end first, so some of the fragment length lists obtained will be in the reverse order. The assembly algorithm can be modified to take this into account.

3) *Applied voltage and current levels*. Blockades are of ionic current flow through the pore due to K+ and Cl- ions in the electrolyte; with V = ~100 mV this current is ~100 pA ($\approx G_{pore}V$, where G_{pore} is the conductance of the pore, typically 1 nS for a pore ~10 nm thick), usually adequate for measuring blockades [6]. With thicker pores blockade levels may be lower. In the presence of noise there is a tradeoff between detectable pulse amplitude changes and translocation speed. While a higher voltage results in a higher blockade current and higher signal-to-noise ratios (SNR), it also causes a fragment or residue with high negative charge to translocate through DNP at a rate that exceeds $1/T_{detector}$, and one with high positive charge to translocate too slowly, resulting in misses or 'loss' to diffusion respectively. These extremes have been estimated in Section A-3 above. (The upper limit to the applied voltage is set by the breakdown field for the electrolyte, typically ~70 MV/m.)

4) *Entropy barriers*. It is assumed that the entropy barrier [6] faced by a fragment during its entry into DNP (*Method 1*) or MNP (*Method 2*) is negligible, in part because short peptides have been considered. Long peptides may form secondary structures and also ball up, impeding entry into a pore. In this case, the barrier may not be negligible; it can be taken into account by increasing the minimum cleaving intervals required of the enzymes. The taper in *trans*2/*cis*3 (Figures 1 and 2) also helps in lowering the entropy barrier. Based on the computational results discussed above, the two methods presented here appear well suited to sequencing of peptides with 12-16 residues. (Compare with the optimum peptide length in an efficient mass spectrometer is ~20 [3].)

5) *Independence of cells*. Each cell targets a different amino acid and operates independent of the other cells. This means that the cell can be independently optimized for enzyme reaction rates, applied voltage, pH value, etc.

6) *Sticky fragments/residues*. The problem of fragments or residues sticking to pore or compartment walls may be resolved through the use of non-stick additives [40] or wall coatings [41].

7) *Sequencing with the potential reversed*. A peptide can be sequenced with the applied potential reversed, which speeds up fragments with positive charge and slows down those with negative charge; neutral fragments are not affected. (If the pore is ion-sensitive, one with the appropriate sense may be used.) Merging the two sets of data can lead to improvements in detection and correction of errors, but this is only for charged fragments. The error can be minimized over all fragments, charged or neutral, by experimentally varying the pH and finding the pH value that yields the best results.

8) *Hafnium oxide pores*. Recent studies using high bandwidth (~4 MHz) detectors have shown that a HfO_2 membrane < 10 nm thick can slow down translocating DNA molecules [42]. (The slowdown is believed to be due to interactions of the DNA with the walls of the pore.) At the present time, however, fabrication seems to require an inordinate amount of time.

9) *Applicability to DNA sequencing*. The counting-based sequencing approach described in the main text can be applied to DNA sequencing if four endonucleases that are distinct and specific to the four nucleotide types can be found or synthesized and can be covalently (or otherwise) attached to the *trans* side of a pore. This could simplify DNA sequencing considerably.

For other implementation-related issues affecting tandem cells see discussions in [11,12].

Additional references

[36] D. L. Nelson and M. M. Cox, *Lehninger's Principles of Biochemistry*, 4th Edition, W. H. Freeman and Company, New York, 2005.
[37] M. W. Germann, T. Turner, and S. A. Allison, "Translational diffusion constants of the amino acids: measurement by NMR and their use in modeling the transport of peptides," *J. Phys. Chem. A*, 2007, **111**, 1452-1455.
[38] R. J. Simpson, *Proteins and Proteomics: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2008.
[39] A. J. Barrett, N. D. Rawlings, and J. F. Woessner, (eds.) *Handbook of Proteolytic Enzymes*, Academic Press, London, 1998.
[40] E. C. Yusko, J. M. Johnson, S. Majd, P. Prangkio, R. C. Rollings, J. Li, J. Yang, and M. Mayer, "Controlling protein translocation through nanopores with bio-inspired fluid walls", *Nature Nanotech.*, 2011, **6**, 253–260.
[41] G. F. Schneider, S. W. Kowalczyk, V. E. Calado, G. Pandraud, H. W. Zandbergen, L. M. K. Vandersypen, and C. Dekker, "DNA translocation through graphene nanopores", *Nano Lett*., 2010, **10**, 3163–3167.
[42] J. Larkin, R. Henley, D. C. Bell, T. Cohen-Karni, J. K. Rosenstein, and M. Wanunu, "Slow DNA transport through nanopores in hafnium oxide membranes," *ACS Nano*, 2013, **7**, 10121–10128.